

Geuvadis analysis group meeting

Genève 2012-04-16

Olof Karlberg & Jonas Almlöf
Uppsala University

RNA concentrations

Qubit conc. (ng/μl)								
QUBIT	Geneve	Barcelona (n-drop)	Berlin	Kiel 1	Kiel 2	Munich	UU	Leiden
HG00117	2260	not analyzed	?	600	572	?	1280	?
HG00355	1440	not analyzed	?	660	600	?	1050	?
NA06986	1640	not analyzed	?	1140	900	?	1490	?
NA19095	1440	not analyzed	?	720	720	?	1280	?
NA20527	1420	1292.91	?	780	660	?	1900	?

BIOANALYZER conc. (ng/μl)							
	Geneve	Barcelona	Berlin	Kiel	Munich	UU	Leiden
HG00117	?	not analyzed	?	not analyzed	?	1460	?
HG00355	?	not analyzed	?	1302	?	1290	?
NA06986	?	not analyzed	?	3534	?	not analyzed	?
NA19095	?	not analyzed	?	1614	?	1440	?
NA20527	?	?	?	1524	?	not analyzed	?

RIN (Bioanalyzer)							
	Geneve	Barcelona	Berlin	Kiel	Munich	UU	Leiden
HG00117	8.9	not analyzed	?	not analyzed	?	8.6	?
HG00355	9.4	not analyzed	?	?	?	9.2	?
NA06986	8.5	not analyzed	?	?	?	not analyzed	?
NA19095	9.4	not analyzed	?	?	?	8.8	?
NA20527	9.3	9.8	?	?	?	not analyzed	?

**Send missing information ASAP to Mathias Brännvall:
mathias.brannvall@medsci.uu.se**

RNA concentrations cont.

- Are the concentrations obtained from the Bioanalyzer good enough
- Can we see concentration effects in the data:
Too high measured concentration => lower complexity due to less input RNA used
- Analysis: correlate concentrations with percentage of duplicates and detection of low abundance transcripts.
Normalize for sequencing depth.

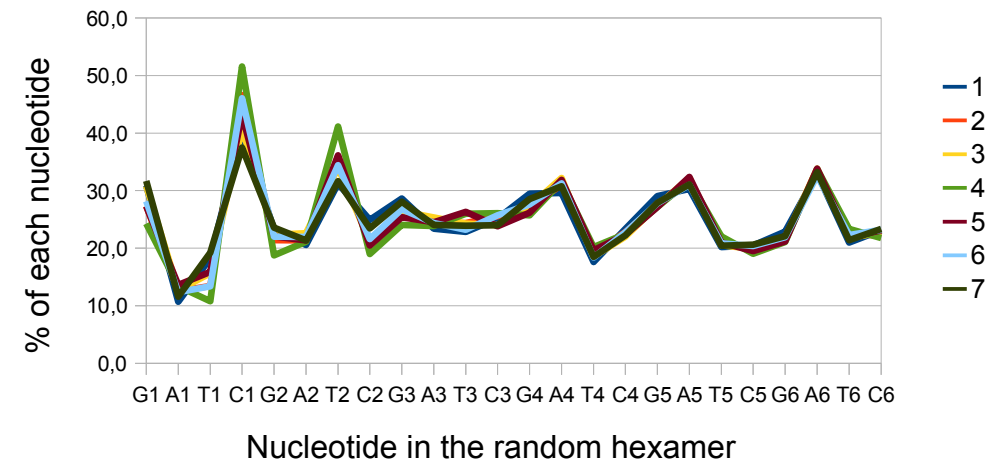
**Send missing information ASAP to Mathias Brännvall:
mathias.brannvall@medsci.uu.se**

Center and sample effects on the random hexamer primer

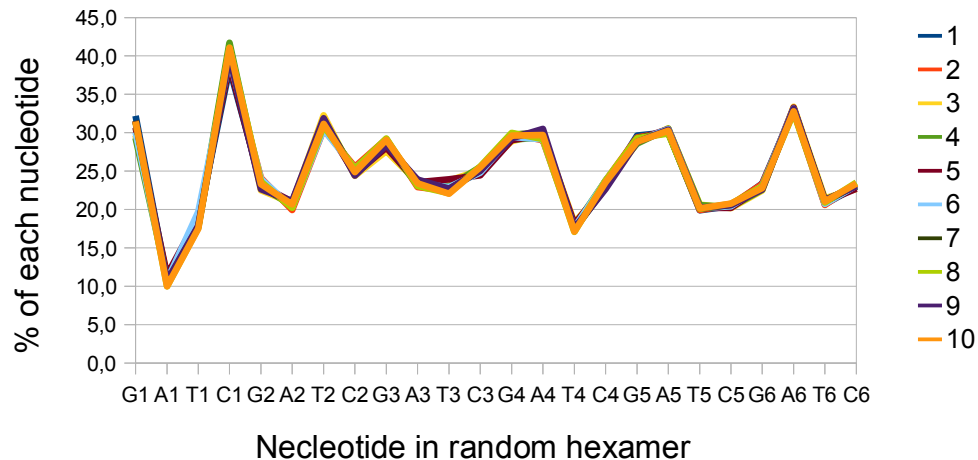
Nucleotide content averaged on sample



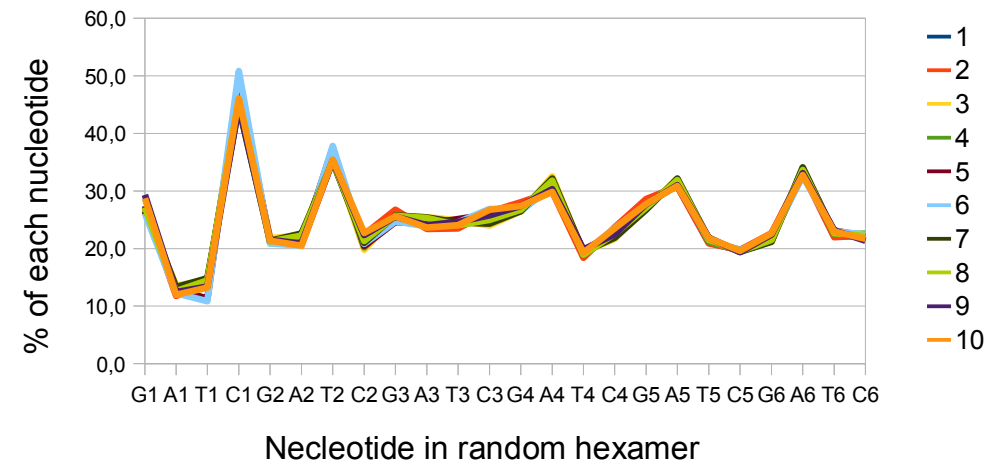
Nucleotide content averaged on center



Nucleotide content in center 1



Nucleotide content in center 2



Investigation of batch effects - outline

- Focus on the 5 common samples
- Clustering of samples using PCA
- Make QC analysis on several different levels
- Coverage smoothing (MART)
- Look for rRNA and DNA content
- Normalization using low coverage data?

Clustering of samples using PCA

- Do the identical samples cluster together tighter than different samples from the same center?
- Variables: GC-content, base content, quality values, coverage conformity, coverage per million reads, data from sample information sheet, ...

QC analysis – data levels

- Raw data (.fastq) – compare data from FastQC and possibly other low level metrics
- Mapped reads (.bam) – Use for example Picards `Collect(RnaSeq|*)Metrics` and/or RSeQC (EVER-seq)
- Expression levels
- If we do not find any center effects on the expression level, does that mean that any bias found at a lower level is irrelevant? Nice to know how much underlying noise is tolerated.

RNA-specific analysis

- Mapped reads distribution
- Coverage uniformity over gene body
- Reproducibility
- Effect of downsampling
- Strand specificity (miRNA)
- Splice junction annotation
- Correct expression/splicing for genotype?

Coverage smoothing

- Remove sequence bias introduced by for example the “random” hexamer primer
- Test effects of smoothing – is it needed?
- Use GLM (poisson linear model) and MART (multiple additive regression trees)
- Is the procedure compatible with current pipeline?

rRNA and DNA content

- How much rRNA is left and does it differ between labs?
- Does the DNA background follow a general poisson distribution or are there outliers? Lab differences?

Normalization using low coverage data

- Use Geneva's low coverage data and sandbox samples to find effect and biases that are either general or center specific
- Correct for biases in all of the samples and see if they make a difference for downstream analysis
- Verify procedure on Geneva's high coverage data

